IT550: Information Retrieval

Assignment 6: Query Expansion and Relevance Feedback

Instructor: Prasenjit Majumder

Learning Outcome: At the end of this assignment you will learn using query expansion and relevance feedback to improve your retrieval.

1 Problem description

In query expansion we extend the initial query by adding related terms to the query. One way would be adding synonyms of words in the query. This can be done using wordnet. In relevance feedback we involve the user in order to improve the retrieval process. Rocchio algorithm is used to perform relevance feedback. In this algorithm we use the equation 1.

$$q_m = \alpha q_0 + \beta \frac{1}{\|D_r\|} \Sigma_{d_j \in D_r} d_j - \gamma \frac{1}{\|D_{nr}\|} \Sigma_{d_j \in D_{nr}} d_j \tag{1}$$

- q_0 is the initial query
- $\|D_r\|$ is the number of relevant documents
- $\|D_{nr}\|$ is the number of non-relevant documents
- q_m is the optimal query

2 Implementation

2.1 Dataset

- For this assignment we will use Telegraph news articles, which is in XML format. It contains news on different categories for the year 2004 to 2007. You can download the dataset from this link: https://drive.google.com/open?id=1JuawXQmYVkjpfL3HOblqjDrqw8V1lHrC
- The Queries are in "en.topics.76-125.2010". The query is of the format shown in Figure 1. Use the sentences enclosed in desc tag for framing your query vector
- The "en.qrels.76-125.2010.txt" contains the documents that are relevant to a query. The format of a qrel is such: Query_No Q0 Document ID Relevance score.
- Relevance score is binary 0 or 1. 1 is for relevant, 0 is for otherwise.
- The documents in the dataset is in the format shown in Figure 2.

2.2 Exercise

1. Implement Rocchio Algorithm for relevance feedback for all the queries.
2. Use $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.15$ as your hyperparameters. Use TF-IDF to represent the documents and queries and perform retrieval. Report the MAP score.
3. Use $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.15$ as your hyperparameters. Use Word2Vec to represent the documents and queries and perform retrieval. Report the MAP score.
4. Perform query expansion on queries using Wordnet. For a word having a synonym attach the synonym to the term. Consider at most 10 synonyms for each word.
5. Use $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.15$ as your hyperparameters. Use TF-IDF to represent the documents and queries and perform retrieval. Report the MAP score.
6. Use $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.15$ as your hyperparameters. Use Word2Vec to represent the documents and queries and perform retrieval. Report the MAP score.

```
<top lang='en'>
<num>76</num>
<title>Clashes between the Gurjars and Meenas</title>
<desc>
Reasons behind the protests by Meena leaders against the
inclusion of Gurjars in the Scheduled Tribes.
</desc>
<narr>
The Gurjars are agitating in order to attain the status of a
Scheduled Tribe. Leaders belonging to the Meena sect have
been vigorously opposing this move. What are the main reasons
behind the Meenas' opposition? A relevant document should
mention the root cause(s) behind the conflict between these
two sects.
</narr>
</top>
```

Figure 1: Query Format

```
<DOC>
<DOCNO> </DOCNO>
<TEXT></TEXT>
</DOC>
```

Figure 2: News Format

3 References

- Tomas Mikolov et al.,Distributed Representations of Words and Phrases and Their Compositionality, NIPS, Volume 2, 2013, 31113119
- TF IDF using sklearn
- Word2Vec using Gensim
- Wordnet for finding synonyms

4 Submission

- You have to submit your assignment in Jupyter notebook with proper comments and explanation of your approach.
- For each of your approach report the MAP score
- The submission deadline for this assignment in **16th March 2020 at 11 PM**